BIOMETRICS 55, 264–267
March 1999

Unconditional Versions of Several Tests Commonly Used in the Analysis of Contingency Tables

Boris Freidlin

Biometric Research Branch, National Cancer Institute,
6130 Executive Blvd. EPN Rm. 739, Bethesda, Maryland 20892-7434, U.S.A.
email: freidlinb@ctep.nci.nhi.gov

and

Joseph L. Gastwirth

Department of Statistics, George Washington University,
Washington, D.C. 20052, U.S.A.

SUMMARY. By focusing on a confidence interval for a nuisance parameter, Berger and Boos (1994, *Journal of the American Statistical Association* **89**, 1012–1016) proposed new unconditional tests. In particular, they showed that, for a 2×2 table, this procedure generally was more powerful than Fisher's exact test. This paper utilizes and extends their approach to obtain unconditional tests for combining several 2×2 tables and testing for trend and homogeneity in a $2 \times K$ table. The unconditional procedures are compared to the conditional ones by reanalyzing some published biomedical data.

KEY WORDS: Cochran–Armitage test; Cochran–Mantel–Haenszel test; Contingency tables; Exact inference; p-value; Unconditional test.

1. Introduction

In many situations where the comparison of the response rates in several groups is of interest, the total number of responses is not fixed. For instance, in a dose–response study using K dose levels, the total number of responses is random. Standard procedures for testing for a trend in the response rates, however, condition on this total, e.g., the implementation of the Cochran–Armitage test in StatXact-3 (Cytel, 1995, p. 479). These conditional procedures are appropriate, but they may not be the most powerful due to the discreteness of the conditional sample space.

For the analysis of 2×2 tables, a variety of unconditional procedures have been developed (for reviews, see Martin and Silva (1994) and Upton (1982)). Suppose the statistic T is used to test the null hypothesis H_0 that the response rates in two groups are equal to an unknown common value π. For the observed value t of T, Barnard (1947) and Suissa and Shuster (1985) took $p_u = \sup_{\pi \in [0,1]}(P_\pi(T \geq t))$ as the p-value of the unconditional test, where π is a nuisance parameter. Mehta and Hilton (1993) showed that using the supremum over all π in $[0, 1]$ entails a conservative p-value. Berger and Boos (1994) considered the supremum of $P_\pi(T \geq t)$ over a $100(1 - \gamma)\%$ confidence region C_γ for π so that the p-value of their unconditional test is given by $p_\gamma = \sup_{\pi \in C_\gamma}(P_\pi(T \geq t)) + \gamma$. Berger (1996) showed that this confidence interval method reduces the conservatism of the unconditional test. We extend the Berger and Boos (1994) approach to obtain unconditional versions of the Cochran–Armitage trend test as well as the test for homogeneity of the response rates in K groups in Section 2. An unconditional combination procedure for several 2×2 tables is given in Section 3.

2. Unconditional Tests for $2 \times K$ Contingency Tables

Consider the response rates in K groups. Let n_i be the sample size of the ith group and X_i the number of responses. Clearly, each X_i is distributed as a binomial random variable with parameters π_i and n_i.

In dose–response or carcinogenicity studies, the alternative of interest is a trend in the π_i. A classic test for trend is the Cochran–Armitage test (Cochran, 1954; Armitage, 1955),

$$T = \sum_{i=1}^{K} v_i (X_i - n_i \hat{\pi}) / \sqrt{V_T}, \tag{1}$$

where v_i are the column scores,

$$V_T = \hat{\pi}(1 - \hat{\pi}) \left[\sum_{i=1}^{K} v_i^2 n_i - \left(\sum_{i=1}^{K} v_i n_i \right)^2 \Big/ \sum_{i=1}^{K} n_i \right],$$

and

$$\hat{\pi} = \sum_{i=1}^{K} x_i \Big/ \sum_{i=1}^{K} n_i.$$

Originally, the asymptotic normality of T was used to analyze large samples. Exact inferences are made using the conditional distribution of T given $\Sigma_{i=1}^{K} x_i$ (cf., Agresti, 1990, p. 118; Cytel, 1995, p. 479).

The unconditional sample space is $\Omega = \{\mathbf{x} = (x_1, x_2, \ldots, x_K) : 0 \le x_i \le n_i,\ i = 1, \ldots, K\}$. Under the null hypothesis H_0: $\pi_1 = \pi_2 = \cdots = \pi_K = \pi$, the probability of each outcome is

$$P_\pi(\mathbf{x}) = \prod_{i=1}^{K} \binom{n_i}{x_i} \pi^{x_i}(1-\pi)^{n_i - x_i}.$$

Suppose outcome $\mathbf{x}$ was observed and $T(\mathbf{x}) = t$. To obtain an unconditional test that extends the approach of Berger and Boos (1994), one first finds a $100(1-\gamma)\%$ confidence region C_γ for π of the form $[\pi_l, \pi_u]$ from the binomial distribution of $Y = \Sigma_{i=1}^{K} X_i$ under H_0, where

$$\pi_l = \sup\left(\pi : \sum_{u \ge y} \binom{N}{u} \pi^u (1-\pi)^{N-u} < \frac{\gamma}{2}\right),$$

$$\pi_u = \inf\left(\pi : \sum_{u \le y} \binom{N}{u} \pi^u (1-\pi)^{N-u} < \frac{\gamma}{2}\right),$$

y is the observed Y, and $N = \Sigma_{i=1}^{K} n_i$. The p-value of the unconditional version of T is

$$p_\gamma = \sup_{\pi \in C_\gamma} (P_\pi(T > t)) + \gamma$$

$$= \sup_{\pi \in C_\gamma} \left(\sum_{(u_1, \ldots, u_K) \in R_T} \prod_{i=1}^{K} \binom{n_i}{u_i} \pi^{u_i}(1-\pi)^{n_i - u_i} \right) + \gamma, \tag{2}$$

where $R_T = \{(u_1, u_2, \ldots, u_K) \in \Omega : T(u_1, u_2, \ldots, u_K) \ge t\}$. Note that the outcomes where all x_i are equal to zero or n_i are placed in the acceptance region, the complement of R_T.

Examples 1 and 2

To illustrate the gain in sensitivity of the unconditional trend procedure, we reanalyze the data from two studies reported in Bickis and Krewski (1989) using equally spaced scores and $\gamma = .001$. In a study of follicular cell adenomas, the following results were reported (number of animals with tumor/number of animals): dose 0, 0/8; dose 0.5, 0/23; dose 1.0, 4/39. The confidence interval (CI) for π is [.0052, .2064]. The one-sided exact p-values of the unconditional and conditional tests are .0474 and .0897, respectively. For a study of occurrence of fibrosarcomas, the data were dose 0, 0/20; dose 0.5, 0/50; dose 1.0, 2/50. The CI for π is [.0003, .0963]. The one-sided exact p-values of the unconditional and conditional tests are .0715 and .1716, respectively. In both cases, the p-value of the unconditional test is about half that of the conditional one.

When the hypothesis of interest is whether the success rates in all K groups are equal, i.e., H_1: $\pi_i \ne \pi_j$ (for at least one pair i and j), the usual statistic (Fleiss, 1981, p. 139) is

$$T_H = \sum_{i=1}^{K} \left(X_i - \frac{n_i m_1}{N}\right)^2 \Big/ \frac{n_i m_1}{N} + \sum_{i=1}^{K} \left(X_i' - \frac{n_i m_2}{N}\right)^2 \Big/ \frac{n_i m_2}{N}, \tag{3}$$

where $X_i' = n_i - X_i$, $m_1 = \Sigma_{i=1}^{K} X_i$, $m_2 = \Sigma_{i=1}^{K} X_i'$, and $N = m_1 + m_2$.

Mehta and Hilton (1993) used T_H to construct an unconditional test for homogeneity based on the unconditional approach of Barnard (1947) and Suissa and Shuster (1985). Using the exact confidence interval for π, the p-value of the unconditional homogeneity test is

$$p_\gamma = \sup_{\pi \in C_\gamma} \left(\sum_{(u_1, \ldots, u_K) \in R_H} \prod_{i=1}^{K} \binom{n_i}{u_i} \pi^{u_i}(1-\pi)^{n_i - u_i} \right) + \gamma, \tag{4}$$

where $R_H = \{(u_1, u_2, \ldots, u_K) \in \Omega_u : T_H(u_1, u_2, \ldots, u_K) \ge T_H(x_1, x_2, \ldots, x_K)\}$.

Example 3

To illustrate the method, we reanalyze, using $\gamma = .001$, the data (Pettriciani, 1985) on infections missed by three early AIDS test kits, i.e., 4/61 (Abbott), 1/92 (Litton), and 1/236 (ENI). The CI for π is [.0025, .0482]. The two-sided exact p-values of the proposed unconditional and conditional (Gastwirth and Johnson, 1989) tests of homogeneity are .0048 and .0073, respectively.

3. An Unconditional Cochran–Mantel–Haenszel Test for Stratified 2 × 2 Tables

Often, the success rates of two treatments are compared over K strata formed by relevant covariates. The number of successes in the ith group of the jth stratum, X_{ij}, is a binomial random variable with parameters π_{ij} and n_{ij} $(i = 1, 2;\ j = 1, \ldots, K)$. We are testing H_0: $\pi_{1j} = \pi_{2j} = \pi_j$ for $j = 1, \ldots, K$ against H_1: $\pi_{1j} \ne \pi_{2j}$ for at least one j. The most common test used for analysis of stratified 2×2 contingency tables is the Cochran–Mantel–Haenszel test (Breslow and Day, 1980, p. 138; Agresti, 1990, p. 230),

$$T_{CMH} = \left[\sum_{j=1}^{K} (X_{1j} - m_{1j} n_{1j}/N_J)\right]^2 \Big/ \sum_{j=1}^{K} \frac{m_{1j} m_{2j} n_{1j} n_{2j}}{(N_i - 1) N_i^2}, \tag{5}$$

where $N_j = n_{1j} + n_{2j}$, $m_{1j} = x_{1j} + x_{2j}$, and $m_{2j} = N_j - m_{1j}$. The exact conditional inference in this setting is based on conditioning on the total number of successes in each stratum (cf., Hirji et al., 1994).

The unconditional sample space is $\Omega_s = \{\mathbf{x} = (x_{11}x_{21}, \ldots, x_{1K}x_{2K}) : 0 \le x_{ij} \le n_{ij}$ for $i = 1, 2;\ j = 1, \ldots, K\}$. Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_j, \ldots, \pi_K)$ denote the vector of stratum specific π_j under the H_0. Then

$$P_{\boldsymbol{\pi}}(\mathbf{x}) = \prod_{j=1}^{K} \binom{n_{1j}}{x_{1j}} \binom{n_{2j}}{x_{2j}} \pi_j^{m_{1j}} (1-\pi_j)^{m_{2j}}.$$

For each stratum j, the $(1-\gamma/K)\%$ confidence interval $C_{\gamma/K}^{j}$ for π_j is calculated. The product C_{s_γ} of the K confidence intervals $C_{\gamma/K}^{j}$ is taken as a conservative joint $100(1-\gamma)\%$

Table 1
Response to thymosin (Li, Simon, and Gart, 1979)

	Group 1		Group 2		Group 3	
	Failure	Success	Failure	Success	Failure	Success
Thymosin	1	10	0	9	0	8
Placebo	1	12	1	11	3	7

confidence interval for the vector π. Suppose the outcome $\mathbf{x}$ was observed and $T(\mathbf{x}) = t$. The p-value of the unconditional Cochran–Mantel–Haenszel test (CMH) is defined as

$$p_\gamma = \sup_{\pi \in C_{s_\gamma}} P_\pi(T_{CMH} \geq t) + \gamma$$

$$= \sup_{\pi \in C_{s_\gamma}} \left(\sum_{(u_{11}, u_{21}, \ldots, u_{1K}, u_{2K}) \in R_{CMH}} \prod_{j=1}^{K} \binom{n_{1j}}{u_{1j}} \times \binom{n_{2j}}{u_{2j}} \pi_j^{m_{1j}} (1-\pi_j)^{m_{2j}} \right) + \gamma,$$

where $R_{CMH} = \{\mathbf{u} = (u_{11}u_{21}, \ldots, u_{1K}u_{2K}) \in \Omega_{us} : T_{CMH}(\mathbf{u}) \geq t\}$, $m_{1j} = u_{1j} + u_{2j}$, and $m_{2j} = N_j - m_{1j}$.

Examples 4 *and* 5

We use the new method, with $\gamma = .001$, to analyze two data sets. Meyskens et al. (1981) reported that the number of relapses in a stratified randomized study of Bacille bilié de Calmette-Guérin (BCG) immunotherapy vs. BCG + high dose of vitamin A in malignant melanomas were, for Stage I, 4 out of 16 for BCG and 1 out of 19 for BCG + high dose of vitamin A and, for Stage II, 2 out of 4 for BCG and 3 out of 10 for BCG + high dose of vitamin A. The one- and two-sided p-values of the proposed exact unconditional test are .0499 and .094, respectively, and the corresponding p-values of the exact conditional test are .090 and .156, respectively. The difference between the p-values is meaningful.

Table 1 presents the data from a study of the response to thymosin in bronchogenic carcinoma patients (Li, Simon, and Gart, 1979). The one-sided p-values of the exact unconditional and conditional tests are .087 and .156, respectively.

4. Monte Carlo Power Study

The power properties of the exact unconditional and conditional tests for trend in $2 \times K$ contingency tables were compared by a simulation study. The successes in each group follow a binomial distribution with parameters (π_i, n_i), where $\pi_i = \exp(\alpha + \beta v_i)/[1 + \exp(\alpha + \beta v_i)]$. Equally spaced scores $\{1, 2, \ldots, K\}$ were used and five balanced 2×3 and 2×4 tables with $n_i = n$ were considered as follows:

Setting 1: 2×3 table, $\pi_1 = .001$, $\pi_2 = .010$, $\pi_3 = .110$ ($\alpha = -9.5$, $\beta = 2.47$) with $n = 20$, 30, 40, and 50,

Setting 2: 2×3 table, $\pi_1 = .025$, $\pi_2 = .076$, $\pi_3 = .206$ ($\alpha = -4.8$, $\beta = 1.15$) with $n = 20$, 30, 40, and 50,

Setting 3: 2×3 table, $\pi_1 = .373$, $\pi_2 = .5$, $\pi_3 = .627$ ($\alpha = -1.04$, $\beta = .52$) with $n = 10$, 20, 30, and 40,

Setting 4: 2×4 table, $\pi_1 = .001$, $\pi_2 = .003$, $\pi_3 = .023$, $\pi_4 = .137$ ($\alpha = -9.5$, $\beta = 1.92$) with $n = 15$, 20, 25, and 30,

Setting 5: 2×4 table, $\pi_1 = .061$, $\pi_2 = .118$, $\pi_3 = .214$, $\pi_4 = .358$ ($\alpha = -3.45$, $\beta = .717$) with $n = 10$, 15, 20, and 25.

Table 2 gives empiric power estimates for the .05-level unconditional and conditional trend tests for the five settings. Each estimate is based on 400 replications. The expected number of responses (ENR) are given in parentheses. When ENR is small (≤ 5), which occurs when the π_i or the sample sizes n_i are small, the power of the unconditional test is greater than that of the conditional one. This improvement in power decreases as the ENR increases. If the ENR exceeds 15, this gain in power is small.

5. Computational Considerations

The unconditional tests were implemented in SAS IML. The algorithm consisted of identifying the set (R_T, R_H, or R_{CMH}) of all outcomes at least as extreme as the observed one. Subroutine NLPDD was then used to find the maximum

Table 2
Empirical power estimates for the five settings. ENR is expected number of responses.

	Sample size (ENR)			
Setting 1				
Test	20 (2.4)	30 (3.6)	40 (4.8)	50 (6.1)
Unconditional	.610	.765	.875	.930
Conditional	.325	.580	.735	.848
Setting 2				
Test	20 (6.1)	30 (9.2)	40 (12.3)	50 (15.4)
Unconditional	.560	.750	.828	.910
Conditional	.465	.678	.793	.878
Setting 3				
Test	10 (15)	20 (30)	30 (45)	40 (60)
Unconditional	.260	.458	.623	.745
Conditional	.213	.458	.623	.745
Setting 4				
Test	15 (2.5)	20 (3.3)	25 (4.1)	30 (4.9)
Unconditional	.638	.733	.855	.895
Conditional	.360	.488	.670	.758
Setting 5				
Test	10 (7.5)	15 (11.3)	20 (15)	25 (18.8)
Unconditional	.538	.728	.833	.915
Conditional	.488	.685	.800	.908

over the nuisance parameter(s). The time required for calculating the p-value depends on the size of the unconditional sample space, e.g., the trend test in a 2×3 table with $n = 20$ (50) took 28 seconds (2 minutes), but a 2×4 table with $n = 50$ took 1.5 hours. For two 2×2 tables with $n_{11} = n_{12} = n_{21} = n_{22} = 20$, CMH took 4 minutes, but if $n_{ij} = 50$, the time required was 2 hours on a Pentium 166 MHz. All the conditional tests were performed using StatXact3.

Acknowledgement

This research was partially supported by a grant from the National Science Foundation. The authors thank the referee and associate editor for helpful comments.

Résumé

En se basant sur un intervalle de confiance du paramètre de nuisance, Berger et Boos (1994, *Journal of the American Statistical Association* **89,** 1012–1016) ont récemment proposé de nouveaux tests non conditionnels. Ils ont notamment montré que, concernant les tables 2×2, cette procédure était en général plus puissante que le test exact de Fisher. Reprenant et prolongeant cette approche, cet article vise à obtenir des tests non conditionnels dans le cas de tables 2×2 stratifiées et de tables $2 \times K$ (tendance ou homogénéité au sein de K groupes). En réanalysant des données médicales déjà publiées, nous comparons ces procédures non conditionnelles aux tests conditionnels correspondants.

References

Agresti, A. (1990). *Categorical Data Analysis.* New York: Wiley.

Armitage, P. (1955). Tests for linear trend in proportions and frequencies. *Biometrics* **11,** 375–386.

Barnard, G. A. (1947). Significance test for 2×2 tables. *Biometrika* **34,** 123–138.

Berger, R. L. (1996). More powerful tests from confidence interval p-values. *American Statistician* **50,** 314–318.

Berger, R. L. and Boos, D. D. (1994). P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* **89,** 1012–1016.

Bickis, M. and Krewski, D. (1989). Statistical issues in the analysis of the long-term carcinogenicity bioassay in small rodents: An empirical evaluation of statistical decision rules. *Fundamental and Applied Toxicology* **12,** 202–221.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research.* Lyon: IARC.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics* **10,** 417–451.

Cytel. (1995). *StatXact-3, User Manual.* Cambridge, Massachusetts: Cytel Software.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions.* New York: Wiley.

Gastwirth, J. L. and Johnson, W. O. (1989). Testing the homogeneity of small error rates: Application to the sensitivity of different Elisa tests for AIDS antibodies. *Statistics and Probability Letters* **7,** 225–228.

Hirji, K. F., Tang, M., Vollset, S. E., and Elashoff, R. M. (1994). Efficient power computation for exact and mid-p test for the common odds ratio in several 2×2 tables. *Statistics in Medicine* **13,** 1539–1549.

Li, S.-H., Simon, M. R., and Gart, J. J. (1979). Small sample properties of the Mantel–Haenszel test. *Biometrika* **66,** 181–183.

Martin Andres, A. and Silva Mato, A. (1994). Choosing the optimal unconditioned test for comparing two independent proportions. *Computational Statistics and Data Analysis* **17,** 555–574.

Mehta, C. R. and Hilton, J. F. (1993). Exact power of conditional and unconditional tests: Going beyond the 2×2 contingency table. *The American Statistician* **47,** 91–98.

Meyskens, F. L., Matti, S. A., Voakes, J. B., Moon, T. E., and Gilmartin, E. (1981). A stratified randomized adjuvant study of BCG + high dose vitamin A in Stage I and II malignant melanoma. In *Adjuvant Therapy of Cancer III,* Salmon and Jones (eds), 217–224. New York: Grune and Stratton.

Pettriciani, J. C. (1985). Licensed tests for antibody of human T-lymphotropic virus type III: Sensitivity and specificity. *Annals of Internal Medicine* **103,** 726–729.

Suissa, S. and Shuster, J. (1985). Exact unconditional sample sizes for the 2×2 binomial trial. *Journal of the Royal Statistical Society, Series A* **148,** 317–327.

Upton, G. J. G. (1982). A comparison of alternative tests for the 2×2 comparative trials. *Journal of Royal Statistical Society, Series A* **145,** 86–105.

Received October 1997. *Revised March* 1998.
Accepted April 1998.